

# 本周周报

解聪(10/28/2013-11/04/2013)

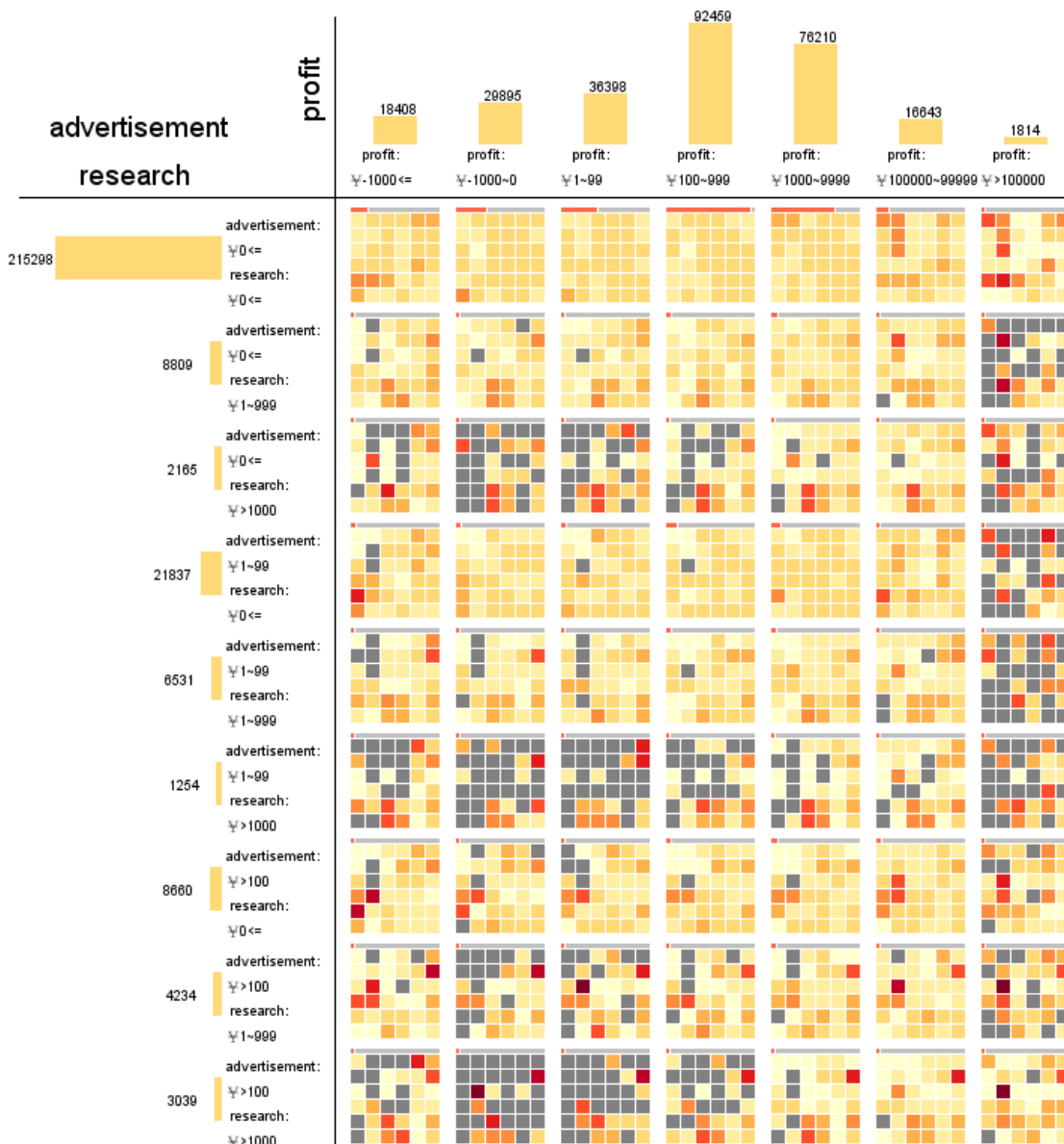
## 一、 组会论文报告

UTOPIAN 是一个交互式主题模型系统。它使用了非负矩阵分解解决了 LDA 主要有两个问题：1.算法多次运行，其输出的结果不一致；2.主题抽取过程中无法结合用户的经验与反馈。

处理过程中，把文档-关键词的对应矩阵看做一个非负矩阵的话，其分解可以帮助我们找到潜在主题。用户可以对非负矩阵的分解进行多种操作：比如主题的合并，主题分裂，以某个文本生成主题，以某个关键词生成主题，修改关键词和主题的对对应关系等等。

## 二、 在 small-mutiple 中探索标签的意义

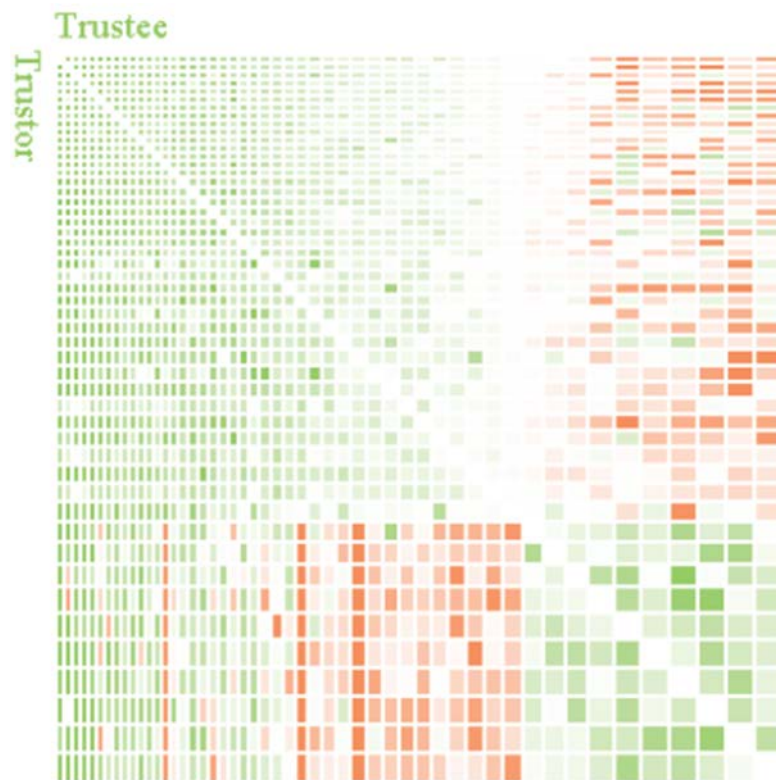
上周对企业数据进行了分析，效果如下，可视编码见上周周报。我们尝试是用投影来分析企业数据之间的关联。采用了 CCA 的方法，这样就得到 7\*9 的矩阵中每两个方块 A 和 B 之间的关联度  $\lambda_{AB}$  (-1 到 1 之间)。



1.投影或者聚类时会使用到两个方块之间的距离，可以使用  $\lambda$  来代替。问题在于，这样的话相关性就不满足距离的三角不等式： $\lambda_{AB} + \lambda_{BC} \geq \lambda_{AC}$ （猜测，未证明）。如果 A 与 B 相似，B 与 C 相似，事实上 A 与 C 可能不相似。导致聚类可能出现这样的矛盾：A 与 B 是一类，B 与 C 一类，但是事实上 A 与 C 并不在一类。

2.又因为我们目的是发现标签与标签的关联，所以是应该吧 small-multiple 中的标签对应的矩阵的每一行或列当做一个整体。重点通过分析行列的关系来分析标签属性的关系。

综合以上两点，考虑由矩阵的操作得到聚类关系。主要就是经过行列的变换重新排列矩阵。有重排之后的矩阵得到标签的关系。按照格子内的相似度，对矩阵的行列进行重排。尽量将相似的行列放置在一起。所以我猜测是：是不是可能得到分组的效果，像 TrustViz 一样？

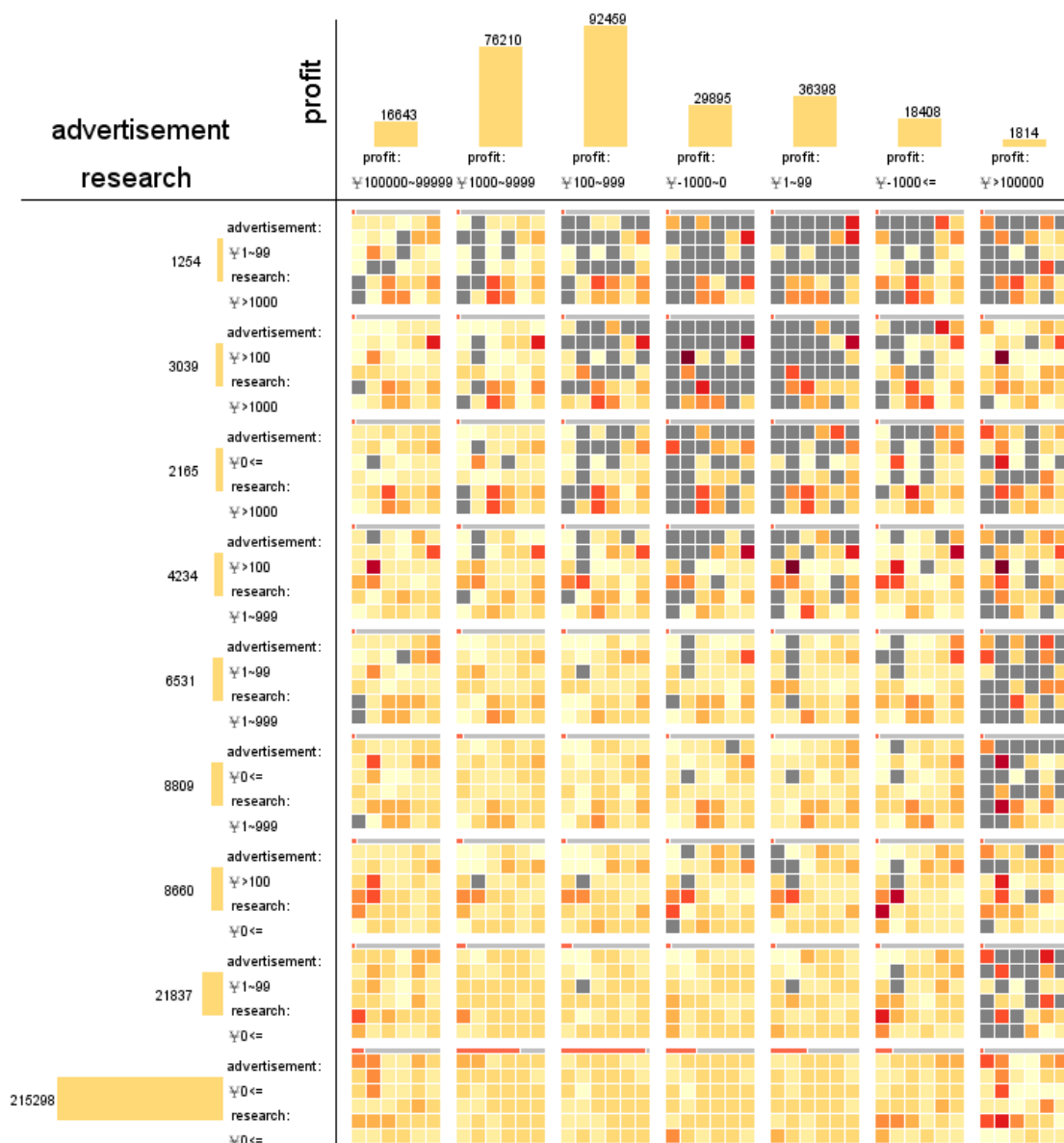


重排后结果如下图。

与上周的结果对比发现：在上周结果中 Y 轴是先按照广告经费分，再按照科研经费分的，是无法看出矩阵在 Y 轴上的规律。用了新的排列以后，矩阵明显在右上角空的比较多（灰色代表空的行业）。再观察 Y 轴的重排方式即可以发现明显的规律：从上到下每一行的科研投入逐渐减小。这说明：

1.科研投入金额越高，这些公司对应的行业数越少。说明很多行业不需要搞科研投入。本方法帮助发现单个维度的一些特征。

2.企业对科研投入的差异大于广告的差异，先选择科研经费划分矩阵优于先使用广告费划分矩阵。使用本方法可以帮助用户确定哪些维度可能有用，那些维度没用，有助发现不同维度的对比。



我个人的理解是：如果做投影我们要考虑更多的问题，如投影结果的准确性，投影与矩阵中每个方块的对应，投影本身不同标签的可视化。我们应该把工作更多地放在 **small-multiple** 的交互（行列变换，单个方块的交互，相似方块的高亮，方块内部 **cell** 的选取等），以及方块的展示上。从而通过 **small-multiple** 本身的可视化与布局发现标签的特征。固定的布局要靠眼睛看出关联，但是很费力；而行列的变换，矩阵的重布局本身就有助于发现不同标签对应的关联。